JAMA | Original Investigation

Evidence of Lack of Treatment Efficacy Derived From Statistically Nonsignificant Results of Randomized Clinical Trials

Thomas Perneger, MD, PhD; Angèle Gayet-Ageron, MD, PhD

IMPORTANCE Many randomized clinical trials yield statistically nonsignificant results. Such results are difficult to interpret within the dominant statistical framework.

OBJECTIVE To estimate the strength of evidence in favor of the null hypothesis of no effect vs the prespecified effectiveness hypothesis among nonsignificant primary outcome results of randomized clinical trials by application of the likelihood ratio.

DESIGN, SETTING, AND PARTICIPANTS Cross-sectional study of statistically nonsignificant results for primary outcomes of randomized clinical trials published in 6 leading general medical journals in 2021.

OUTCOME MEASURES The likelihood ratio for the null hypothesis of no effect vs the effectiveness hypothesis stated in the trial protocol (alternate hypothesis). The likelihood ratio quantifies the support that the data provide to one hypothesis vs the other.

RESULTS In 130 articles that reported 169 statistically nonsignificant results for primary outcomes, 15 results (8.9%) favored the alternate hypothesis (likelihood ratio, <1), and 154 (91.1%) favored the null hypothesis of no effect (likelihood ratio, >1). For 117 (69.2%), the likelihood ratio exceeded 10; for 88 (52.1%), it exceeded 100; and for 50 (29.6%), it exceeded 1000. Likelihood ratios were only weakly correlated with *P* values (Spearman *r*, 0.16; *P* = .045).

CONCLUSIONS A large proportion of statistically nonsignificant primary outcome results of randomized clinical trials provided strong support for the hypothesis of no effect vs the alternate hypothesis of clinical efficacy stated a priori. Reporting the likelihood ratio may improve the interpretation of clinical trials, particularly when observed differences in the primary outcome are statistically nonsignificant.

JAMA. 2023;329(23):2050-2056. doi:10.1001/jama.2023.8549

Editorial page 2023

Supplemental content

Author Affiliations: Division of Clinical Epidemiology, Geneva University Hospitals, and Faculty of Medicine, University of Geneva, Geneva, Switzerland.

Corresponding Author: Thomas Perneger, MD, PhD, Division of clinical epidemiology, Geneva University Hospitals, Boulevard de la Tour 8, 1211 Geneva 14, Switzerland (thomas.perneger@hcuge.ch).

The purpose of randomized clinical trials is to provide solid and actionable evidence about the effectiveness of new treatments. When a well-conducted randomized trial finds a statistically significant benefit of a new treatment, the interpretation of the evidence is straightforward: the data indicate that the new treatment works. But when the between-group difference is not statistically significant, which happens frequently,[1-4] evidence is seen as inconclusive. Researchers are cautioned against confusing lack of statistical significance with lack of effect, as captured in the saying "absence of evidence is not evidence of absence."[5,6] This reasoning follows the fisherian statistical tradition: an experiment can only provide evidence *against* the tested hypothesis, which is never proven correct.[7] Thus researchers who obtained a statistically nonsignificant result are stuck: they cannot claim effectiveness any more than they can deny it. This explains the proliferation of noncommittal conclusions such as, "The treatment did not significantly improve outcomes." Such sayings are mere repetitions of the observed result—as if a clinician told a patient their test was negative, without counseling them about the presence or absence of disease.

One way of addressing this issue is to consider the whole CI for the measure of treatment efficacy, such as a risk difference, odds ratio, or hazard ratio [HR].[3,4,8] Any parameter value within the CI is deemed compatible with the observed data.[9] If the CI of a statistically nonsignificant result excludes clinically meaningful values, one might conclude that no clinically meaningful effect exists; however, CIs typically include both clinically meaningful and clinically negligible values of the effectiveness parameter, and such results remain ambiguous.[8,10]

This study proposes an alternative approach, based on a direct comparison of the null hypothesis of no effect to the alternate hypothesis of effectiveness (as stated by the researchers in their sample size calculation) in light of observed trial data. The reasoning resembles the diagnostic process, when a clinician weighs the probabilities of a disease being absent or present in light of the result of a diagnostic test. The comparison is expressed in terms of a likelihood ratio (LR),[11,12] which quantifies the strength of support provided by the observed results to one hypothesis vs the other. We apply this method to a sample of published statistically nonsignificant clinical trial

findings to estimate how strongly such results support the null hypothesis over the alternate.

Key Points

Question Can a statistically nonsignificant result of a randomized clinical trial provide conclusive evidence of lack of effect of the new treatment?

Findings Among 169 statistically nonsignificant primary outcome results of randomized trials published in 2021, the hypotheses of lack of effect (null hypothesis) and of clinically meaningful effectiveness (alternate hypothesis) were compared using a likelihood ratio to quantify the strength of support the observed trial findings provide for one hypothesis vs the other; about half (52.1%) yielded a likelihood ratio of more than 100 for the null hypothesis of lack of effect vs the alternate.

Meaning Many statistically nonsignificant clinical trial results demonstrate conclusive evidence of lack of effect of the new treatment.

Methods

Selection of Articles

We conducted a cross-sectional study of statistically nonsignificant results in reports of randomized clinical trials published in the *Annals of Internal Medicine*, *BMJ*, *JAMA*, *Lancet*, *New England Journal of Medicine*, and *PLoS Medicine* from January 1, 2021, through December 31, 2021 (**Figure 1**). We selected these journals because they publish quality research across a broad spectrum of topics. We searched PubMed with filters for the publication dates, *randomized clinical trial* as article type, and *journal name* as source. We included articles that reported a statistically nonsignificant result (2-tailed *P* value >.05 or 95% CI that included the null value) for a primary or coprimary outcome from a phase 3 superiority trial, with a CI for the measure of effectiveness, and that stated the alternate hypothesis, ie, the magnitude of the treatment effect used for the sample size calculation. Statistically nonsignificant results for secondary or exploratory outcomes were not included.

Because we analyzed data in the public domain, we did not seek approval from an institutional review board.

Variables

The main variables were the treatment effect, its CI, the *P* value, and the alternate hypothesis. Descriptive variables were type of trial (2-group, multigroup, factorial), type of intervention (drug or biologic substance, dosage or mode of delivery, nonpharmacological intervention), comparator (placebo or sham intervention, specified comparison treatment, usual care), and scale of measure of effect (multiplicative, such as HRs, or additive, such as difference in means).

Data Collection

The primary treatment effect and CI were retrieved from the abstract, and from the Results section of each article. The alternate hypothesis was retrieved from the sample size calculation specified in the Methods section (or when this information was not found in the article, we searched prior publications or the published trial protocol). When necessary, we transformed the alternate hypothesis into the metric used to report the treatment effect; eg, when proportions (p_O and p_1) were used for sample size calculations, we transformed them into a relative risk (RR) (p_1/p_O), odds ratio $[p_1(1-p_O)/p_O(1-p_1)]$, or relative hazard $[\log(1-p_1)/\log(1-p_O)]$, as needed.

The initial data collection was performed by the first author and verified by the second author. Discrepancies were resolved by consultation with the source documents.

Computation and Interpretation of the LR

For each statistically nonsignificant trial result, we computed an LR for the null hypothesis (H_N) vs the alternate hypothesis (H_A) specified by the investigator (**Box**). The LR measures the relative support given by the data to one hypothesis over another; eg, an LR of 5 means that the data support the null hypothesis 5 times as strongly as the alternate hypothesis.[11,12] It is a symmetrical measure: LR = 5 for H_N vs H_A is equivalent to LR = 1/5 for H_A vs H_N.

The LR is independent of prior beliefs about the null and alternate hypotheses. Posterior odds for one hypothesis over the other equal a product of prior odds (ie, prior beliefs) by the LR, according to the Bayes theorem. For illustration, we show posterior probabilities of the null hypothesis (vs the alternate) as a function of prior probabilities and of the observed LR (**Table 1**). Interpretation guidelines suggest that an LR of 10 represents strong evidence, an LR of 100 represents decisive evidence,[14] and an LR of 100 or more lead to high posttest probabilities that the null hypothesis is true rather than the specific alternate hypothesis for common levels of prior belief.

Statistical Analysis

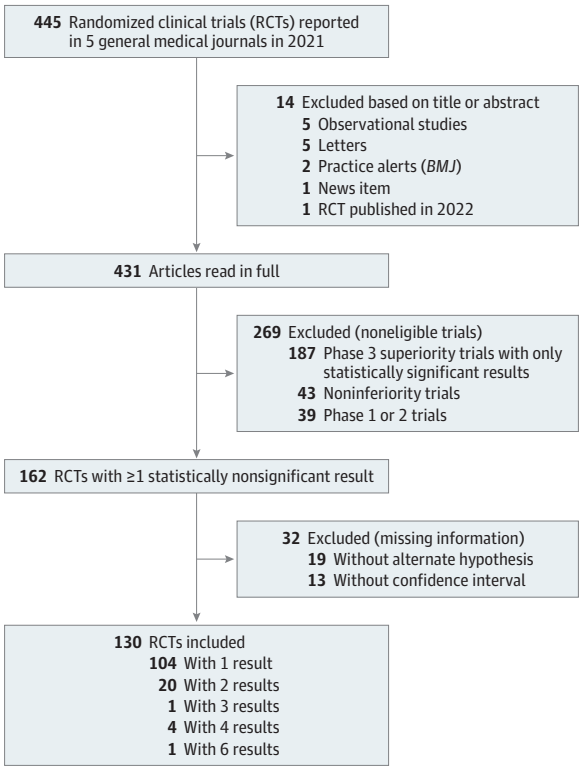
We describe frequency distributions for descriptive categorical variables. LRs for the null hypothesis vs the alternate are described in natural units or are categorized at approximate half-powers of 10 (ie, 1, 3, 10, 30...). *P* values were grouped in 4 categories with thresholds of .25, .50, and .75 (when a *P* value was not reported, we computed an approximate *P* value from the *z* statistic). We examined proportions of results with LRs exceeding 100 across descriptive categorical variables. We also determined whether the effectiveness parameter under the alternate hypothesis was within the observed 95% CI.

We also provide a scatterplot of LRs vs corresponding *P* values and computed a Spearman correlation coefficient. Analyses were conducted using SPSS version 25.

Results

The search retrieved 431 reports of randomized clinical trials published in the selected journals in 2021 (Figure 1), among which 162 included at least 1 statistically nonsignificant result for a primary outcome. No CIs or no specific alternate hypothesis in 32 trials were found, so the 130 articles that were included (eAppendix in Supplement 1) reported 169 statistically nonsignificant results for primary outcomes. Of note, 1 trial of vaccine efficacy[15] sought to reject a nonnull hypothesis (vaccine efficacy of 25%)

Figure 1. Flow Diagram of Study Selection



Box. Computation of the Likelihood Ratio

The likelihood ratio corresponds to the ratio of the probability densities (f) of the observed result (eg, the z statistic for the primary outcome) under the 2 hypotheses[11]

$$LR = \frac{f_N(z)}{f_A(z)}$$

The likelihood ratio can be easily computed from published trial results.[12] When z is normally distributed, the natural logarithm of the likelihood ratio for the null hypothesis vs the alternate hypothesis (H_N vs H_A) is

$$\ln(LR) = z(N - A) - \frac{1}{2}(N^2 - A^2)$$

Here z is the test statistic for the measure of effect (ie, $\hat{\theta}/se[\hat{\theta}]$, where $\hat{\theta}$ is the estimate of the effectiveness parameter obtained in the trial, as $se[\hat{\theta}]$ its standard error). N is the value of the effectiveness parameter under H_N expressed in $se(\hat{\theta})$ units (ie, $\theta_N/se[\hat{\theta}]$). In most cases N equals 0. A is the value of the effectiveness parameter under H_A expressed in $se(\hat{\theta})$ units (ie, $\theta_A/se[\hat{\theta}]$). The effectiveness parameter θ is left in natural units for differences of means or proportions and was logarithm transformed for multiplicative measures, such as the hazard ratio. The standard error $se(\hat{\theta})$ can be computed from the 95% CI for θ (or $\log[\theta]$), by dividing the width of the CI by 3.92 (2 × 1.96).

An online calculator[13] is available for the computation of the likelihood ratio from the value of the effectiveness parameter under the alternate hypothesis θ_A and from the point estimate $\hat{\theta}$ and its 95% CI: https://medresearch.shinyapps.io/Bayesian_re-analysis/. The likelihood ratio is computed for the alternate hypothesis vs the null; take the inverse to get the likelihood ratio for the null vs the alternate.

rather than the null (vaccine efficacy of 0%). Another trial[16] specified 2 alternate hypotheses for each of 2 comparisons—a reduction of the outcome event and a symmetrical increase; this situation was treated as separate comparisons of hypotheses, and LRs for the 4 contrasts were obtained.

The complete study database is available in Supplement 2.

Characteristics of Included Results

The largest set of results was published in *JAMA*, followed by the *New England Journal of Medicine* and *PLoS Medicine* (**Table 2**). Most results came from 2-group trials. A majority of the experimental interventions were substances (drugs, supplements, biologics, vaccines) and their delivery methods or dose variations. The rest were nonpharmacological interventions (device, surgery, diagnostic method, behavioral intervention, exercise, health care organization). Control treatments were roughly evenly divided between usual care, specific active interventions, and placebo or sham interventions. More results were expressed on a multiplicative scale (45 relative hazards, 38 RRs, 23 odds ratios, 2 geometric mean ratios, 2 incidence rate ratios, 1 ratio of RRs) than on a linear scale (36 differences in means, 20 risk differences). *P* values ranged from .052 to >.99, with 25th, 50th, and 75th quartiles of .19, .44, and .67.

Likelihood Ratios

The median LR value was 115.6 (range, 0.2-10[146]; IQR, 5.1-2661); deciles were 1.1, 3.0, 8.3, 33.1, 115.6, 302, 982, 54 113, and 9 041 714. Fifteen LRs (8.9%) were less than 1; 154 (91.1%) exceeded 1; 117 (69.2%) exceeded 10; 88 (52.1%) exceeded 100; and 50 (29.6%) exceeded 1000 (**Table 3**).

The trial findings with the lowest and highest LRs, and examples of those with values near 1, 10, 100, 1000, and 10 000, are shown in **Table 4**. The lowest value of 0.2 was obtained from a trial of fetal surgery vs expectant care among infants with diaphragmatic hernia.[17] The RR of survival to discharge from intensive care was 1.27 (95% CI, 0.99-1.63). The LR favors the alternate hypothesis (fetal surgery is superior to expectant management) 5-fold over the null hypothesis (no difference); the observed result was closer to the effect specified by the alternate hypothesis (RR, 1.36) than to the null value of 1.0. When a trial's observed result lay midway between the hypothesized null and alternate effects,[18] the LR was near unity, reflecting roughly equivalent support for the 2 hypotheses. In a trial comparing hemodynamic-guided management with usual care for patients with heart failure,[19] a quantitatively similar HR and CI yielded an LR of 9.6 in favor of the null hypothesis because the stated alternate hypothesis specified a greater efficacy (HR, 0.70 vs 0.79). Larger LRs in favor of the null[20-23] reflected observed results that were close to absence of effect, with CIs that were ever further away from the parameter value under the alternate hypothesis of effectiveness.

About half the LRs for the null hypothesis over the alternate exceeded 100, a proportion that was somewhat lower for

Table 1. Posttest Probability of the Null Hypothesis as a Function of the Pretest Probability (3 Levels) and Likelihood Ratio (5 Levels)[a]

Likelihood ratio for the null hypothesis vs the alternate hypothesis[b]	Pretest probability (odds) of the null hypothesis of ineffectiveness being true vs the alternate hypothesis of effectiveness[c]		
	Equipoise, 50% (1:1)	Cautious optimism, 25% (1:3)	Strong optimism, 10% (1:9)
0.1 (favors the alternate)	9.1	3.2	1.1
1	50	25	10
10	90.9	76.9	52.6
100	99.0	97.1	91.7
1000	99.9	99.7	99.1

[a] All values are percents. The posttest probability of the treatment being ineffective (rather than effective) is computed using the Bayes theorem. Likelihood ratios of 100 or more lead to high posttest probabilities of the null hypothesis being true, rather than the specific alternate hypothesis, at all 3 common levels of prior belief.

[b] The likelihood ratio represents the relative support given by the data to hypotheses of ineffectiveness (null hypothesis) vs effectiveness (alternate hypothesis).

[c] The pretest probability represents the researchers' pretrial beliefs about the new treatment being ineffective rather than effective at a clinically relevant level stated in the protocol.

Table 2. Trial Characteristics

Characteristic[a]	No. (%)		Proportion with LR >100 in favor of the hypothesis of ineffectiveness, %
	Articles	Results	
Total	130 (100)	169 (100)	52.1
Journal			
JAMA	41 (31.5)	54 (32.0)	59.3
New England Journal of Medicine	26 (20.0)	33 (19.5)	45.5
PLoS Medicine	21 (16.2)	32 (18.9)	46.9
BMJ	21 (16.2)	23 (13.6)	47.8
Lancet	13 (10.0)	19 (11.2)	63.2
Annals of Internal Medicine	8 (6.2)	8 (4.7)	37.5
Type of trial			
2 Groups	105 (80.8)	118 (69.8)	50.0
Multigroup	15 (11.5)	30 (17.8)	56.7
Factorial	10 (7.7)	21 (12.4)	57.1
Type of intervention			
Substance[b]	62 (47.7)	76 (45.0)	51.3
Delivery or dose	12 (12.3)	17 (10.1)	58.8
Nonpharmacological[c]	52 (40.0)	76 (45.0)	51.3
Type of comparator			
Usual care	49 (37.7)	67 (39.6)	46.3
Active control	45 (34.6)	54 (32.0)	59.3
Placebo or sham	36 (27.7)	48 (28.4)	52.1
Scale of measure of effect			
Multiplicative (eg, hazard ratio)	86 (66.2)	113 (66.9)	45.1
Additive (eg, difference in means)	44 (33.8)	56 (33.1)	66.1
P value (24 calculated)			
.05 to <.25		53 (31.4)	45.3
.25 to <.50		43 (25.4)	55.8
.50 to <.75		45 (26.6)	57.8
.75 to >.99		28 (16.6)	50.0

Abbreviation: LR, likelihood ratio.

[a] Characteristics of 130 randomized trials with 169 statistically nonsignificant results published in 2021.

[b] Drugs, supplements, biologics, vaccines.

[c] Devices, surgery, diagnostic methods, behavioral interventions, exercise, health care delivery, or organization.

multiplicative measures of effect than for additive measures but otherwise did not vary much by type of trial (Table 2).

LRs vs *P* Values

The scatterplot of LRs vs *P* values showed only a weak association (Spearman correlation coefficient, 0.16; *P* = .045; **Figure 2**).

LRs vs CIs

Thirty-nine CIs (23.1%) included the value of the effectiveness parameter under the alternate hypothesis; the remainder (130, 76.9%) did not. When the 95% CI included parameter values under both hypotheses, the LRs ranged from 0.2 to 6.2. When it did not, LRs ranged from 3.0 to 10^{146}.

Table 3. Distribution of Likelihood Ratios for the Null Hypothesis vs the Alternate Hypothesis

Likelihood ratio[a]	No. (%) (n = 169)
>0.1 to 1	15 (8.9)
>1 to 3	18 (10.7)
>3 to 10	19 (11.2)
>10 to 30	13 (7.7)
>30 to 100	16 (9.5)
>100 to 300	20 (11.8)
>300 to 1000	18 (10.7)
≥1000	50 (29.6)

[a] Likelihood ratio categories represent approximate half-powers of 10 (eg, 1, 3, 10, 30, 100) for 130 trials with 169 statistically nonsignificant results for primary outcomes published in 2021. Likelihood ratios less than 1 represent support for the alternate hypothesis; those that are equal to 1 represent equal support for the 2 hypotheses; and those that exceed 100 represent strong support for the null hypothesis.

Discussion

This study revealed that many—but not all—statistically nonsignificant results from randomized clinical trials provide compelling evidence in favor of the hypothesis of no effect vs the hypothesis of effectiveness specified in the trial protocol. Based on a sample of 169 statistically nonsignificant primary outcome findings published in leading medical journals, roughly 70% of LRs for the null vs the prespecified alternate hypothesis exceeded 10, 50% exceeded 100, and 30% exceeded 1000. If researchers were in equipoise at the start of the trial, LRs of more than 100 would increase the posttrial probability that the treatment is ineffective rather than effective at the prespecified level, to levels approaching certainty (see the Box). In such cases, further research on the treatment should likely stop. The remaining results (≈30%) provided weak to moderate support for either hypothesis, which would justify further research on the topic. Statistically nonsignificant trial results may lead to opposite conclusions regarding lack of effectiveness and to different courses of further action, depending on the magnitude of the LR for the null hypothesis vs the alternate.

These observations rely on the application of the LR to the trial hypotheses, which is not a common procedure at present. The LR employs the same data summary as the statistical test, ie, the *z* statistic but with a different purpose: the test derives a *P* value as evidence against the null hypothesis, whereas the LR compares the relative support that data provide to the 2 competing hypotheses of the trial. Although larger *P* values were associated with larger LRs for the null hypothesis, the rank correlation was modest, and large or small LRs were observed at any given *P* value (Figure 2). Clearly the information content of the 2 procedures differs.

Current statistical guidelines indicate that statistically nonsignificant results lack evidential content.[5,6] Researchers are tempted to use "spin" to make their results more palatable[24] or to look for explanations for this undesirable outcome.[25] Yet the analysis of LRs shows that for at least half of such results, the observed data strongly support the hypothesis of no effectiveness over the prespecified hypothesis of effectiveness. This contradiction stems from different definitions of statistical evidence. Under the *P*-value paradigm, evidence—represented by small *P* values—can only be obtained against the tested hypothesis (the null); thus, when the *P* value is large, evidence is deemed "absent." There is no concept of evidence in favor of the tested hypothesis, and the alternate hypothesis does not even feature in the reasoning. (The limitations of the *P* value as measure of evidence have been addressed.[11,26]) Under the LR paradigm, only comparative evidence can be obtained for a pair of hypotheses—ie, the relative support data provide for one hypothesis over the other. The 2 hypotheses are treated evenly. There is no concept of absolute evidence for or against a hypothesis; only comparisons are possible.[11,27] The statement "absence of evidence is not evidence of absence" remains logically true, but the LR finds usable evidence where the significance test saw none. The evidence derived from statistically nonsignificant trial results conveys important information for regulators, investigators, and clinicians, which is missed when only statistical significance is considered.

LRs also improve the interpretation of CIs for the treatment effect. CIs are easy to interpret when they contain only clinically meaningful or only clinically negligible values of the effectiveness parameter. But often they contain some of both,[10] in which case they do not provide usable evidence. LRs are not similarly limited. Furthermore, in this study, that a CI included the null parameter value but not the alternate did not imply that the null hypothesis was well supported because LRs in favor of the null ranged from about 3 to exponentially large values. Computation of the LR remains valuable for an accurate interpretation of such results.

Limitations

This study has several limitations. First, the sample used in this study was not necessarily representative of all statistically nonsignificant trial results; it is an illustrative data set, based on publications in reputable journals in a single year. A second limitation is the choice of a relevant alternate hypothesis. We chose to accept at face value the investigators' definition of a meaningful treatment effect, stated in their protocols. However, investigators may be tempted to overestimate this effect so as to make the required sample size fit resource constraints. Moderately favorable trial results would yield strong support for the null hypothesis if the alternate hypothesis was overly optimistic. However, users of evidence are not held to the investigators' choices and can obtain an LR for the null hypothesis vs the alternate of their choosing by using the formula given in this article or using a calculator available online.[13]

Third, the concept of the LR may be objectionable to those who feel uncomfortable considering only 2 simple hypotheses, represented by 2 numerical values of the parameter of interest, as though other parameter values were impossible. The 2 hypotheses do not exhaust all possibilities and the full CI should be considered as well. However, the 2 hypotheses represent paradigmatic situations stated a priori, one in which the experimental treatment doesn't work, the other in which it has clinically important efficacy. Comparing their merits

Table 4. Examples of 2021 Trials at Various Levels of Support for the Null Hypothesis Over the Alternate Hypothesis of Effectiveness[a]

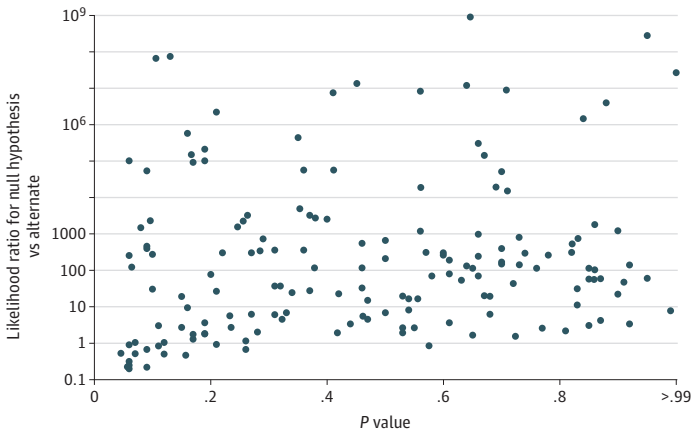
Source	LR for null hypothesis vs alternate	Population	Intervention		Primary outcome	Alternate hypothesis	Result (95% CI)	LR (exact)[b]
			Experimental	Comparator				
Deprest et al[17]	Lowest	Fetuses with diaphragmatic hernia	Fetal surgery	Expectant care	Survival to discharge from intensive care	RR, 1.36 (75% vs 55%, from protocol)	RR, 1.27 (0.99 to 1.63)	0.20
Meyerhardt et al[18]	Near 1	Patients with stage III colon cancer	Celecoxib (and standard therapy)	Placebo (and standard therapy)	Disease-free survival	HR, 0.79	HR, 0.89 (0.76 to 1.03)	1.05
Lindenfeld et al[19]	Near 10	Patients with heart failure	Hemodynamic-guided management	Usual care	Composite event	HR, 0.70	HR, 0.88 (0.74 to 1.05)	9.6
Jiménez et al[20]	Near 100	Patients hospitalized with COPD	Diagnostic strategy for pulmonary embolism	Usual care	Composite event	RD, −10%	RD, 0.5% (−6.2% to 7.3%)	103.3
Hopewell et al[21]	Near 1000	Patients with rotator cuff disorders	Progressive exercise	Single-session advice	Shoulder pain and disability index	Difference in means, −8	Difference, −0.66 (−4.52 to 3.20)	981.7
NIHR Global Research[22]	Near 10 000	Surgery patients	Chlorhexidine disinfection	Iodine disinfection	Surgical site infection	RR, 0.67	RR, 0.97 (0.82 to 1.14)	15 132
Okereke et al[23]	Highest	Adults without depression	Supplementation with omega-3 fatty acids	Placebo	PHQ-8 depression scale	Difference in means, −0.5	Difference, 0.03 (−0.01 to 0.07)	10^{146}

Abbreviations: COPD, chronic obstructive pulmonary disease; HR, hazard ratio; LR, likelihood ratio; PHQ-8, 8-Item Patient Health Questionnaire; RD, risk difference; RR, risk ratio.

[a] See Supplements 1 and 2 for a complete listing of all trial sources, settings, design, findings, and likelihood ratio calculation variables.

[b] The likelihood ratio for the trial's null hypothesis vs alternate hypothesis was calculated based on the observed treatment effect, accompanying CI, and alternative hypothesis specified in the reported sample size calculation or protocol. Likelihood ratio values less than 1 favor the alternate hypothesis; values that exceed 1 favor the null hypothesis.

Figure 2. Scatterplot of Likelihood Ratios and *P* Values of Statistically Nonsignificant Results



The likelihood ratio for the null hypothesis vs alternate hypothesis as a function of the corresponding *P* value for 160 statistically nonsignificant results for primary outcomes in randomized clinical trials published in 2021 (9 likelihood ratios >10^9 are not shown). The y-axis is on a logarithm scale. Spearman correlation coefficient = 0.16 (*P* = .045).

in light of data are thus legitimate. Some authors would rather move to a bayesian analysis of clinical trial results.[28] This is an appealing option, but one that poses challenges of its own. The main obstacle is the dependence on a prior distribution for the parameter of interest. The subjective nature of the prior distribution is considered an undesirable feature by many scientists.[29] Furthermore, the goal of bayesian analysis is to describe how the observation of data modifies beliefs, from prior to posterior[26-28]; its focus is not on evidence as such. In contrast, the LR isolates what the data say about the relative merits of 2 hypotheses independent of the personal beliefs of the researchers. Two persons who hold different prior beliefs about the effectiveness of the new treatment will end up with different posterior beliefs once the trial results become available—as they should, per the Bayes theorem—even though they use the same evidence represented by the LR.

Conclusions

In this study of statistically nonsignificant primary outcome results of randomized clinical trials published in 2021, a large proportion of results provided strong support for the hypothesis of no effect vs the alternate hypothesis of clinical efficacy stated a priori by the researchers. Reporting the LR may improve the interpretation of clinical trials, particularly when observed differences in the primary outcome are statistically nonsignificant.

ARTICLE INFORMATION

Accepted for Publication: May 1, 2023.

Author Contributions: Dr Perneger had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: All authors.
Acquisition, analysis, or interpretation of data: All authors.
Drafting of the manuscript: Perneger.
Critical revision of the manuscript for important intellectual content: Gayet-Ageron.
Statistical analysis: Perneger.
Administrative, technical, or material support: All authors.

Conflict of Interest Disclosures: None reported.

Data Sharing Statement: See Supplement 3.

Additional Contributions: We thank Philip M. Jones, MD, MSc, University of Western Ontario, London, Ontario, Canada, who created the LR calculator.[13]

REFERENCES

1. Djulbegovic B, Lacevic M, Cantor A, et al. The uncertainty principle and industry-sponsored research. *Lancet*. 2000;356(9230):635-638. doi:10.1016/S0140-6736(00)02605-2

2. Cho D, Roncolato FT, Man J, et al. Clinical equipoise for trials of novel biologic therapies, therapeutic success rates, and predictors of success: a meta-analysis. *JCO Precis Oncol*. 2017;1:1-12. doi:10.1200/PO.17.00062

3. Gewandter JS, McDermott MP, Kitt RA, et al. Interpretation of CIs in clinical trials with non-significant results: systematic review and recommendations. *BMJ Open*. 2017;7(7):e017288. doi:10.1136/bmjopen-2017-017288

4. Gates S, Ealing E. Reporting and interpretation of results from clinical trials that did not claim a treatment difference: survey of four general medical journals. *BMJ Open*. 2019;9(9):e024785. doi:10.1136/bmjopen-2018-024785

5. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ*. 1995;311(7003):485. doi:10.1136/bmj.311.7003.485

6. Alderson P. Absence of evidence is not evidence of absence. *BMJ*. 2004;328(7438):476-477. doi:10.1136/bmj.328.7438.476

7. Fisher RA. *The Design of Experiments*. 1st ed. Oliver and Boyd Ltd; 1935:19.

8. Hemming K, Javid I, Taljaard M. A review of high impact journals found that misinterpretation of non-statistically significant results was common. *J Clin Epidemiol*. 2022;145:112-120. doi:10.1016/j.jclinepi.2022.01.014

9. Gelman A, Greenland S. Are confidence intervals better termed "uncertainty intervals"? *BMJ*. 2019;366:l5381. doi:10.1136/bmj.l5381

10. Perneger TV, Brindel P, Combescure C, Gayet-Ageron A. Evidence of survival benefit was often ambiguous in randomized trials of cancer treatments. *J Clin Epidemiol*. 2020;127:1-8. doi:10.1016/j.jclinepi.2020.06.026

11. Royall RM. *Statistical Evidence—A Likelihood Paradigm*. Chapman & Hall; 1997.

12. Perneger TV. How to use likelihood ratios to interpret evidence from randomized trials. *J Clin Epidemiol*. 2021;136:235-242. doi:10.1016/j.jclinepi.2021.04.010

13. Jones PM. Bayesian re-analysis of biomedical research. Version 1.0. Accessed June 2, 2023. https://medresearch.shinyapps.io/Bayesian_re-analysis/

14. Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc*. 1995;90:773-795. doi:10.1080/01621459.1995.10476572

15. Gray GE, Bekker LG, Laher F, et al; HVTN 702 Study Team. Vaccine efficacy of ALVAC-HIV and bivalent subtype C gp120-MF59 in adults. *N Engl J Med*. 2021;384(12):1089-1100. doi:10.1056/NEJMoa2031499

16. Albert CM, Cook NR, Pester J, et al. Effect of marine omega-3 fatty acid and vitamin D supplementation on incident atrial fibrillation: a randomized clinical trial. *JAMA*. 2021;325(11):1061-1073. doi:10.1001/jama.2021.1489

17. Deprest JA, Benachi A, Gratacos E, et al; TOTAL Trial for Moderate Hypoplasia Investigators. Randomized trial of fetal surgery for moderate left diaphragmatic hernia. *N Engl J Med*. 2021;385(2):119-129. doi:10.1056/NEJMoa2026983

18. Meyerhardt JA, Shi Q, Fuchs CS, et al. Effect of celecoxib vs placebo added to standard adjuvant therapy on disease-free survival among patients with stage III colon cancer: the CALG/SWOG 80702 (Alliance) randomized clinical trial. *JAMA*. 2021;325(13):1277-1286. doi:10.1001/jama.2021.2454

19. Lindenfeld J, Zile MR, Desai AS, et al. Haemodynamic-guided management of heart failure (GUIDE-HF): a randomised controlled trial. *Lancet*. 2021;398(10304):991-1001. doi:10.1016/S0140-6736(21)01754-2

20. Jiménez D, Agustí A, Tabernero E, et al; SLICE Trial Group. Effect of a pulmonary embolism diagnostic strategy on clinical outcomes in patients hospitalized for COPD exacerbation: a randomized clinical trial. *JAMA*. 2021;326(13):1277-1285. doi:10.1001/jama.2021.14846

21. Hopewell S, Keene DJ, Marian IR, et al; GRASP Trial Group. Progressive exercise compared with best practice advice, with or without corticosteroid injection, for the treatment of patients with rotator cuff disorders (GRASP): a multicentre, pragmatic, 2 × 2 factorial, randomised controlled trial. *Lancet*. 2021;398(10298):416-428. doi:10.1016/S0140-6736(21)00846-1

22. NIHR Global Research Health Unit on Global Surgery. Reducing surgical site infections in low-income and middle-income countries (FALCON): a pragmatic, multicentre, stratified, randomised controlled trial. *Lancet*. 2021;398(10312):1687-1699. doi:10.1016/S0140-6736(21)01548-8

23. Okereke OI, Vyas CM, Mischoulon D, et al. Effect of long-term supplementation with marine omega-3 fatty acids vs placebo on risk of depression or clinically relevant depressive symptoms and on change in mood scores: a randomized clinical trial. *JAMA*. 2021;326(23):2385-2394. doi:10.1001/jama.2021.21187

24. Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*. 2010;303(20):2058-2064. doi:10.1001/jama.2010.651

25. Pocock SJ, Stone GW. The primary outcome fails—what next? *N Engl J Med*. 2016;375(9):861-870. doi:10.1056/NEJMra1510064

26. Berger JO, Sellke T. Testing a point null hypothesis: the irreconcilability of *P* values and evidence. *J Am Stat Assoc*. 1987;82(26):112-122. doi:10.2307/2289131

27. Royall RM. The likelihood paradigm for statistical evidence. In: Taper ML, Lele SR, eds. *The Nature of Scientific Evidence*. University of Chicago Press; 2004:119-138. doi:10.7208/chicago/9780226789583.003.0005

28. Spiegelhalter DJ, Friedman LS, Parmar MKB. Bayesian approaches to randomized trials. *J R Stat Soc Ser A Stat Soc*. 1994;157:357-387. doi:10.2307/2983527

29. Efron B. Why isn't everyone a Bayesian? *Am Stat*. 1986;40(1):1-5.